DOI https://doi.org/10.15407/csc.2023.01.065
UDC 004.822

V.V. ZOSIMOV, Doctor of Technical Sciences, Professor, Department of Applied, Information Systems, Taras Shevchenko National University of Kyiv, Bohdan Hawrylyshyn, str. 24, Kyiv, 04116, Ukraine, Scopus Author ID 57188682230, ORCID: https://orcid .org/ 0000-0003-0824-4168, zosimovvv@gmail.com

ENHANCING ONLINE SEARCH SECURITY THROUGH BAYESIAN TRUST NETWORK IMPLEMENTATION

The article focuses on the development of an information search and protection system based on a Bayesian trust network as a measure of document relevance to the user's query. The result is the development of search system structures and algorithms with relevance evaluation when searching the Internet, the implementation of data transmission with an adaptive database for storing decisions. If the need arises, when the goal set before the user cannot be achieved without involving additional information, the adaptive database sends a request to the search system, which in turn collects the necessary information. Mathematical formalization of the definition of relevant decisions was carried out. A graph was modelled, which was built based on Bayesian Trust Networks (BTN) in the GeNIe application package.

***Keywords**: Bayesian Trust Networks, Internet search, Relevance ranking, Query processing, online privacy.*

Introduction

In today's era of rapidly growing internet and the proliferation of information sources, finding relevant information has become increasingly challenging. The sheer volume of data available online makes it difficult for traditional search algorithms to effectively navigate and deliver accurate results. This is where Bayesian networks come into play, offering a powerful solution for improving the efficiency of information search on the internet. Despite their effectiveness, the implementation of Bayesian networks in real-world applications has been limited thus far, with much room for further development and refinement. In this article, we aim to address this issue by exploring the potential of Bayesian networks for improving information search in the internet and highlighting current best practices in their usage [1].

The availability of vast amounts of information on the internet has led to the increasing need for effective information search systems. The traditional search engines, while being useful, have limitations in accurately providing relevant information to users. In recent years, Bayesian Networks (BNs) have been gaining popularity as a tool for improving the effectiveness of information search. BNs are probabilistic graphical models that represent the relationships between variables and the uncertainties associated with them. They are well suited for information retrieval as they can capture the probabilistic dependencies between query and documents, as well as the uncertainty of the user's information need [2].

Development of a Mathematical Model of the System

Let there be a certain number of documents that have been obtained from the Internet network.

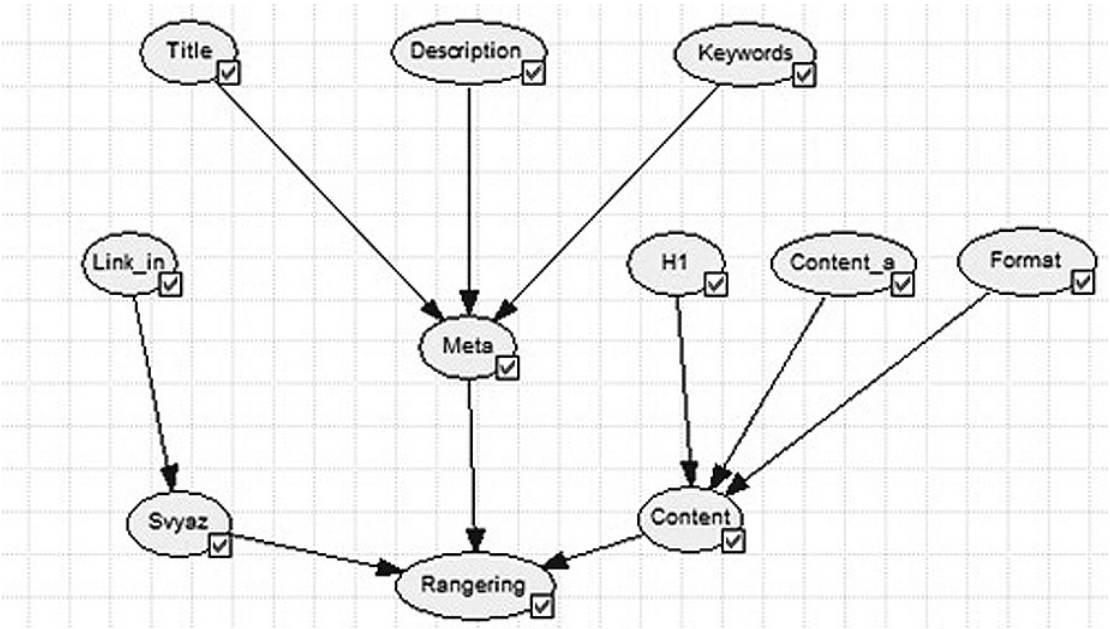

Fig. 1. Bayesian network for calculating the relevance of decision making

Each document has certain characteristics with various attributes that are inherent to certain documents. In the search system, the required keywords are specified that characterize the information that needs to be found. The set of keywords and criteria is called a search query [3].

The degree of correspondence of each specific document to the query is called the relevance of the document to the query. The task of the search system is to present the user with the most relevant results that match his query as much as possible.

Therefore, it is necessary to construct an intellectual system that allows to assess the degree of relevance of each document to the given query. The degree of relevance is a number, based on its value, documents can be sorted by relevance. This number should have the following parameters:

- a non-negative real number;
- the degree of relevance will be higher, the higher the relevance itself of the document to the query;
- the degree of relevance should be expressed in a quantitative form that allows to perform mathematical operations with it.

A Bayesian network is used for modeling domains that are characterized by uncertainty. This uncertainty can be due to a lack of understanding of the domain or a combination of given factors.

Practical Implementation

The following factors that influence the relevance of a search document were identified in the implementation of the system [4]: The factors listed in Table 1 are represented as nodes in a Bayesian network, each of which can take corresponding states and are given tables of conditional probabilities for these nodes. Upon receipt of a search query, the system performs a calculation of each factor for each keyword and carries out the propagation of the corresponding calculated probabilities in the network. The result of the work is the probability $P(C \mid F_1, F_2 ... F_n)$ for each available document D, which is the measure of the relevance of the document to the query [5].

If $P(C = c_1 \mid F_1, F_2 ... F_n) > P(C = c_2 \mid F_1, F_2 ... F_n)$ then the document is relevant to the query, that is

$P(C = c_1 \mid F_1, F_2 ... F_n) > 0.5)$, then $D \in C_1$. (1)

Documents that satisfy the decision rule (1) are passed to the dynamic subsystem with a standardized measure of relevance.

Table. **Factors Included as Nodes in the Bayesian Network**

Factor	Distinguishing states	Explanation
Inclusion of a keyword	Inclusions	The presence of a keyword in the position
Inclusion of a keyword in the document header.	1) 0 inclusions 2) 1 and more inclusions	The presence of a keyword in the position title increases the relevance of the searched document; absence reduces relevance
Inclusion of the keyword in the brief description of the document	1) 0 inclusions 2) Exactly 1 inclusion	The presence of a keyword in the summary (the first 25 words of the article) increases the relevance of the article; absence does not change relevance
Multiple occurrences of the keyword in the brief description of the document	1) Less than 2 inclusions 2) 2 and more inclusions	The inclusion of the keyword in the summary two or more times increases the relevance of the article
The number of inclusions of the keyword in the searched text	1) Less than 2 inclusions 2) From 2 to 7 inclusions 3) More than 7 inclusions	The presence of a keyword in the text 2 or more times increases the relevance of the query (non-linearly, according to the discrete values of the factor "2", "3", "4", "5", "6", "7 and more"); the presence of exactly one occurrence does not change the relevance, the absence of occurrences reduces the relevance
The position of the keyword in the text of the document	The value ranges from 0.6086 to 1	The position of the word in the text of the document is represented by a number from 0 (the end of the document) to 1 (the beginning of the document); a larger value of this number increases the relevance (non-linearly, by continuous values of the factor)
The number of occurrences of **bigrams** (pairs of words) in the title of the document	1) 0 inclusions 1. 2) 1 and more inclusions	The presence in the position title of a phrase (in bigrams) that coincides with a phrase of two keywords increases relevance; the absence of a phrase does not change the relevance
The number of occurrences of **bigrams in the full text** of the search document	1) 0 inclusions 2) From 1 to 4 inclusions 3) more than 4 inclusions	The presence of bigrams in the full text of the article (summary + text + skills) by 1 or more times increases the relevance of the document (non-linearly, according to the discrete values of the factor "1", "2", "3", "4 and more"); the absence of a phrase does not change the relevance of the document
The **TF*IDF** factor value for the keyword	1) 0 2) Values range from 0 to 4 2) A value greater than 4	A higher value of the factor TF * IDF [6], which takes into account the frequency of occurrence of the keyword (TF) and the weight of the word in the document (IDF), increases the relevance of the document (non-linearly, according to continuous values of the factor in the interval from 0 to 4, with the value "4 and more" - maximum); a value of 0 does not change the relevance of the document
Date of publication of the article	Values in the interval from 0 to 50 (days)	The factor represents the number of days that have passed since the publication of the article to the current (today's) date. A higher value of the factor reduces the relevance of the document (non-linearly, according to discrete values "1", "2", ... "49", "50 and more"); the value 0 ("today") does not change the relevance of the article

Node Name	State Name	Special Name F...	Special Name	Node Id	State Id	Prior Probability	Cost	Type	Ranked	Mandatory	Target State	De...	No...	Qu...	St...	Tr...	Links
Content				Node11				Auxiliary					Add				Add
	True														Add		Add
	False														Add		Add
Content_a				Node8				Auxiliary					Add				Add
	True					0.28									Add		Add
	False					0.72									Add		Add
Description				Node4				Auxiliary					Add				Add
	True					0.65823									Add		Add
	False					0.34177									Add		Add
Format				Node15				Auxiliary					Add				Add
	True					0.31									Add		Add
	False					0.69									Add		Add
H1				Node7				Auxiliary					Add				Add
	True					0.36									Add		Add
	False					0.64									Add		Add
Keywords				Node5				Auxiliary					Add				Add
	True					0.54									Add		Add
	False					0.46									Add		Add
Link_in				Node9				Auxiliary					Add				Add
	True					0.44									Add		Add
	False					0.56									Add		Add
Meta				Node6				Auxiliary					Add				Add
	True														Add		Add
	False														Add		Add
Rangering				Node13				Auxiliary					Add				Add
	True														Add		Add
	False														Add		Add
Svyaz				Node12				Auxiliary					Add				Add
	True														Add		Add
	False														Add		Add
Title				Node3				Auxiliary					Add				Add
	True					0.72									Add		Add
	False					0.28									Add		Add

Fig.2. Filling in the source data

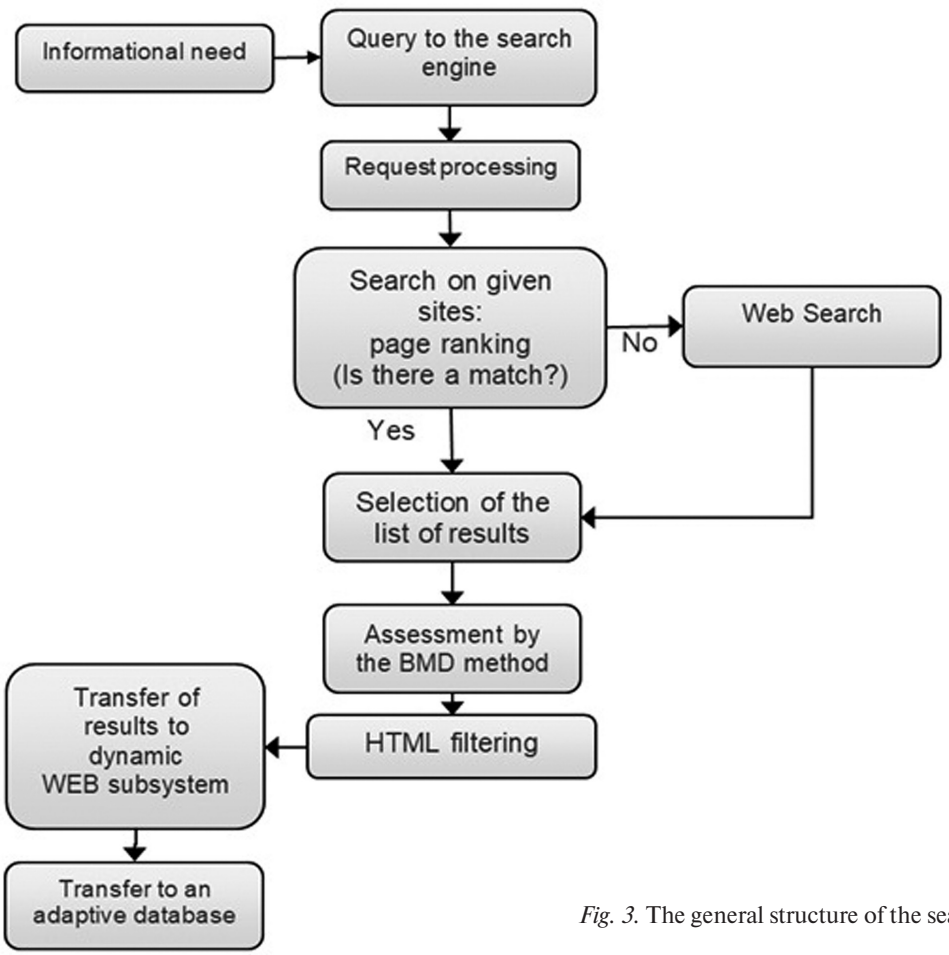

Fig. 3. The general structure of the search engine

$$P = \frac{P(C \mid F_1, F_2 ... F_n) - P_{\min}}{P_{\max} - P_{\min}} \cdot 100\% = \\ = 2 \cdot [P(C \mid F_1, F_2 ... F_n) - 0.5] \cdot 100\%. \quad (2)$$

Modelling the Operation of the System on the Genie 2.0 PPZ

Based on the calculations, it is possible to construct a generalized graph for determining the probability of the normalized measure of relevance of a search query, which is shown in Fig. 1. Modeling work in the Genie 2.0 program is shown on Fig.2.

Determination of the probability begins with the entry of information into the network on nodes of external factors and the results of ranking on nodes of personal factors. Let the values shown in Fig. 2 be known at the time of starting the calculations.

Ranking Function

The ranking function for search is one of the first ways to implement a probabilistic search model, but it already uses several features of the document and query, such as the length of the document and query in words, the frequency of the query text in the document, etc. This function was built manually: the specific formula and numerical values of the parameters were obtained by the authors based on the semantics of the used features [7].

Many authors propose a large number of query and document features that may be useful for ranking, such as the ratio of incoming and outgoing links of the document, the length of the URL, the age of the document, and many others. Any model of user behavior actually adds another feature for the ranking function. Modern search systems use several hundred different features for ranking, so manual construction of the ranking function is simply impossible.

Today, the most popular method for building a ranking function is based on machine learning. Pairs (query, document) are input into the system as vectors of their features, and a training set is formed based on expert assessments. The rankings of the function are automatically built based on this information and a certain type of machine learning method. Depending on the type of assessments provided by the experts, several approaches are distinguished.

1. Pointwise. Each (query-document) pair is assigned its own assessment. If the assessment is categorical, the classification task arises it is necessary to predict the class for (query-document) pairs; if the assessment is numerical, the regression task.

2. Pairwise. For document pairs corresponding to one query, experts provide "pairwise preferences" which document is more relevant to the query.

The General Structure of the System

The structure of the developed search system is shown in Fig. 3.

General principle of operation:

- there is an information need, which comes in the form of a signal from an adaptive database. It arises in the case of an insufficient knowledge base and for the further search for alternative answers;
- the search engine itself reacts to the incoming signal, after which it starts;
- after starting the search engine, the request is sent for special processing: checking for errors, which are checked with a regular expression;
- then the search is carried out on previously specified web pages. With the help of ranking, the search engine checks any textual content on the page;
- if no matches are found, the search engine turns to general Internet resources and conducts a second search there;
- the next block evaluates the results using the Bayesian trust network method;
- in the future, the results are recorded in html format and transferred to the dynamic web subsystem;
- after the specified operations, the dynamic web subsystem transfers the result to the adaptive database.

Protection of user's personal data in the information search process:

The developed system implements the function of blocking the use of cookies, which ensures the anonymity of the user.

With this feature, the user's browser does not leave footprints on websites. Accordingly, the user will be protected from the use of cookies, tracking

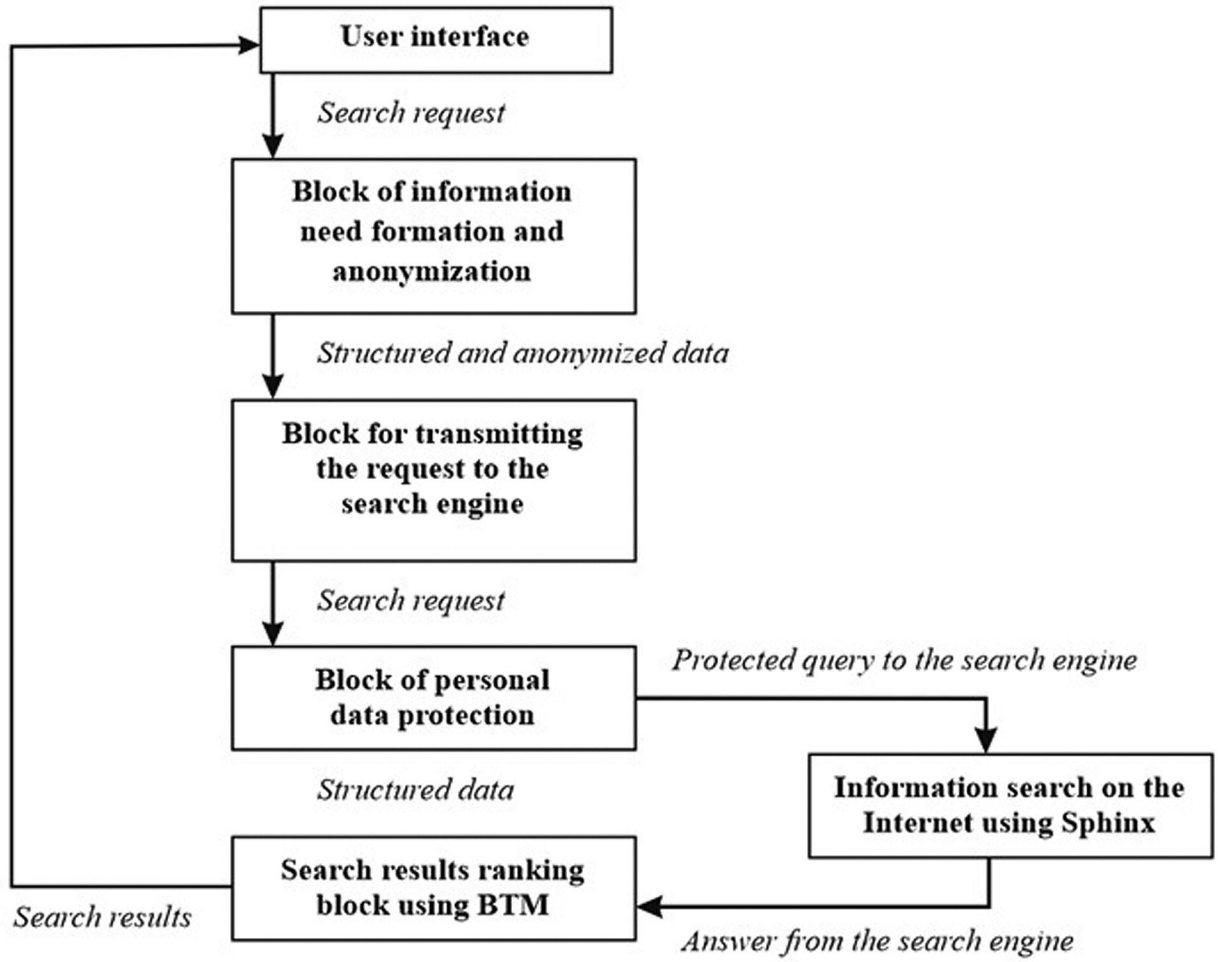

Fig. 4. The algorithm of the developed secure data search system in the Internet

pixels of social networks and other privacy interventions while browsing the website.

Unnecessary metadata, including IP address and other user-identifying information, is automatically removed from a user's request during a search. The anonymized search query is then sent to the server, which returns the search results [8].

The algorithm of the developed search system consists of such main stages:

1. Ensuring the protection of the user's personal data by anonymizing the search request.

2. Carrying out an initial search for information on the Internet using the Sphinx search engine.

3. Ranking and additional processing of primary search results using implemented algorithms using the Bayesian trust network.

Fig. 4 presents the flowchart of the algorithm's operation.

Conclusion

In this paper, the authors present the development of an information search and protection system based on Bayesian network of trust. The proposed system aims to improve the relevance of document retrieval by incorporating trust-based relationships between the user and the information sources. The system is designed to address the limitations of traditional search engines by providing more accurate results based on the user's needs and trust in information sources. The authors have formalized the

relevance determination process mathematically and have modeled the trust network using a Bayesian network. The proposed system has been implemented and tested, showing promising results in improving the relevance of information retrieval.

In the field of information retrieval, the use of BNs has gained traction as a means of improving the effectiveness of information search. The proposed system is a valuable contribution to the field as it demonstrates the potential of using BNs for trust-based information search and protection. The results of the study provide insights into the practical applications of BNs for information retrieval and the potential for further research in this area.

REFERENCES

1. *Gallego, C.*, 2018. "A review of Bayesian deep learning techniques and their application to computer vision prob-lems". Big Data Analytics, IGI Global, pp. 11–25.
2. *Guo, C.*, 2017. "Deep Bayesian active learning for neural networks". Journal of Machine Learning Research, Vol. 18, pp. 1–47.
3. *Pelt, M.,* 2019. "Uncertainty quantification in deep learning using Bayesian convolutional neural networks". Journal of Computer Vision. Vol. 126, pp. 617–635.
4. *Sattari, P.*, 2020. "Bayesian deep reinforcement learning: A survey". Journal of Machine Learning Research, Vol. 21, pp. 1–35.
5. *Zosimov, V., Bulgakova, O., Pozdeev, V.,* 2021. "Complex internet data management system". Advances in Intelligent Systems and Computing. Vol.1246, pp. 639–652.
6. *Zosimov, V., Bulgakova, O.*, 2020. "Calculation the Measure of Expert Opinions Consistency Based on Social Profile Using Inductive Algorithms". Advances in Intelligent Systems and Computing. Vol. 1020, pp. 622–636.
7. *Nalisnick, M.*, 2019. "Deep Bayesian neural networks with many irrelevant inputs". Proceedings of the 35th In-ternational Conference on Machine Learning. Vol. 97, pp. 1748–1757.
8. *Hron, J.*, 2018. "Probabilistic programming for deep learning: A review". Machine Learning Research. Vol. 19, pp. 1–41.

Received 12.02.2023

В.В. Зосімов, доктор технічних наук, професор, Київський
національний університет України імені Тараса Шевченка,
04116, м. Київ, вул. Богдана Гаврилишина, 24, Україна, Scopus Author ID 57188682230,
ORCID: https://orcid .org/ 0000-0003-0824-4168,
zosimovvv@gmail.com

ПІДВИЩЕННЯ БЕЗПЕКИ ПОШУКУ ІНФОРМАЦІЇ В МЕРЕЖІ ІНТЕРНЕТ ШЛЯХОМ ВПРОВАДЖЕННЯ БАЙЄСІВСЬКОЇ МЕРЕЖІ ДОВІРИ

Вступ. У сучасну епоху швидкого зростання Інтернету та поширення джерел інформації пошук відповідної інформації стає дедалі складнішим. Величезний обсяг даних, доступних в Інтернеті, ускладнює ефективну навігацію та надання точних результатів традиційним пошуковим алгоритмам. Байєсовські мережі пропонують потужне рішення для підвищення ефективності пошуку інформації в Інтернеті. Попри на їхню ефективність, реалізація байєсівських мереж у реальних додатках наразі була обмеженою, з великим простором для подальшого розвитку та вдосконалення.

Мета статті. Метою статті є дослідити потенціал байєсівських мереж для покращення пошуку інформації в Інтернеті та висвітлити сучасні кращі практики використання цих мереж. Побудова на основі отриманих результатів дослідження прототип системи підвищення безпеки пошуку інформації в мережі Інтернет.

Методи. Системний підхід, аналіз.

Результати. Запропонована система має на меті покращити релевантність пошуку документів завдяки забезпеченню довірчих відносин між користувачем та джерелами інформації. Систему розроблено для усунення обмежень традиційних пошукових систем, надаючи точніші результати на основі потреб користувача та довіри

до джерел інформації. Автори математично формалізували процес визначення релевантності та змоделювали довірчу мережу за допомогою байєсівської мережі. Запропоновану систему було впроваджено та випробувано, й вона показала багатонадійні результати щодо підвищення актуальності пошуку інформації.

Висновки. У сфері інформаційного пошуку використання *BN* набуло поширення як засіб підвищення ефективності пошуку інформації. Запропонована система є цінним внеском у цю сферу, оскільки вона демонструє потенціал використання *BN* для пошуку та захисту інформації на основі довіри. Результати дослідження дають змогу зрозуміти практичне застосування *BN* для пошуку інформації та потенціал для подальших досліджень у цій галузі.

Ключові слова: *Байєсовські мережі довіри, пошук в Інтернеті, рейтинг релевантності, обробка запитів, конфіденційність в Інтернеті.*